TEC-0070

Task-Driven Active Vision for Security and Surveillance

H. Keith Nishihara
Phil Kahn
Stanley J. Rosenschein
J. Brian Burns
Matthew Turk

Teleos Research
576 Middlefield Road
Palo Alto, CA 94301

January 1996

Approved for public release; distribution is unlimited.

Prepared for:
Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington, VA 22203-1714

Monitored by:
U.S. Army Corps of Engineers
Topographic Engineering Center
7701 Telegraph Road
Alexandria, Virginia 22315-3864

US Army Corps
of Engineers
Topographic
Engineering Center

Destroy this report when no longer needed.
Do not return it to the originator.

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

The citation in this report of trade names of commercially available products does not constitute official endorsement or approval of the use of such products.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** January 1996	**3. REPORT TYPE AND DATES COVERED** Annual Technical Oct. 1993 - Oct. 1994

4. TITLE AND SUBTITLE

Task-Driven Active Vision for Security and Surveillance

5. FUNDING NUMBERS

DACA76-93-C-0017

6. AUTHOR(S)

H. Keith Nishihara   Stanley J. Rosenschein   Phil Kahn
J. Brian Burns   Matthew Turk

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Teleos Research
576 Middlefield Road
Palo Alto, CA 94301

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Advanced Research Projects Agency
3701 North Fairfax Drive, Arlington, VA 22203-1714

U.S. Army Topographic Engineering Center
7701 Telegraph Road, Alexandria, VA 22315-3864

19. SPONSORING / MONITORING AGENCY REPORT NUMBER

TEC-0070

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Teleos is developing and integrating active vision technology with intelligent, domain-specific control strategies to provide real-time visual detection, characterization, and tracking of objects in the security and surveillance domain. Work to date has focused on theoretical studies and experimentation in the areas of attentional mechanisms, figure-ground discrimination, recognition, and active sensor mechanisms to support visual surveillance. The primary research focus is on the detection and tracking of humans in a monitored area. This report describes new results in blob-based shape recognition; the use of oriented texture blobs for face-finding; and stereo based real-time figure-ground discrimination techniques.

14. SUBJECT TERMS

Physical Security, Recognition, Figure-Ground,
Real-Time Tracking

15. NUMBER OF PAGES

36

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** Unlimited

Contents

List of Figures

Preface

This research is sponsored by the Advanced Research Projects Agency (ARPA) and monitered by the U.S. Army Topographic Engineering Center (TEC), under Contract DACA76-93-C-0017 titled, "Task-Driven Active Vision for Security and Surveillance." The ARPA point of contact is Dr. Pradeep Khosla and the TEC Contracting officer's Representatives are Ms. Lauretta Williams and Mr. Thomas Hay.

1 Introduction

A key goal of Teleos' efforts has been the study and implementation of practical, task-driven computer vision. Our view is that the sophisticated performance observed in biological systems is to a large degree derived from the fluent use of simple and robust measurement capabilities. The task being performed drives how and when sensory actions are to be performed. A key element of this research strategy is the tight integration of perception, action selection, and action execution. For example, experienced drivers apply strategies for controlling where and when to monitor the position of the roadway, signs, and other objects; inexperienced drivers are noted for their poor attentional control strategies. Such attentional controls are also a key element of effective surveillance and security tasks. In the first year of this project, Teleos has developed key technologies required for task-driven active vision systems: visual measurement primitives and visual attentional control mechanisms.

Modular real-time active vision measurement capabilities can be applied to a large set of task-oriented systems. Hard perception problems are easier to solve with a good set of primitive measurement capabilities. To be of practical value, a visual measurement capability must provide robust and appropriate measurements in time to be useful at a cost that is not prohibitive. One consequence of these considerations has been an increased focus in our research program on demand-driven visual measurements as opposed to the more traditional approach, which attempts to carry out a full scene analysis prior to making any use of the information derived. The traditional approach has the disadvantage of being computationally expensive and does not easily support the differing needs of diverse applications. Instead of doing full image processing in isolation, our approach is to concentrate on making basic local measurements well.

A visually rich, but otherwise restricted, application domain is vital for guiding and evaluating a core research effort of this type. Surveillance and Security (S&S) provides a rich research context in which to test task-directed vision techniques. It supports perceptual tasks of interestingly different types over a broad range of difficulty. These include detection of new or reappearing objects, classification of movement patterns, detection of common destinations, detection and tracking of motion in visual or IR imagery with an active head, and discrimination of humans using size, shape, color, texture, and motion cues. S&S problems often require visual strategies in order to perform well, and goal-directed attentional mechanisms are key in all but simple cases. Finally, surveillance and security has a wide range of important governmental and commercial applications that include: law-enforcement and security (e.g., detection of public criminal activities such as drug dealing on street corners, building surveillance, detection of loitering and parking lot security), nuclear storage warehouse security, and consumer mobility pattern analysis in a store (e.g., to detect shoplifters, optimize product placement for consumer traffic patterns).

This Annual Report details progress that Teleos has made in the development of computer vision and visual attention mechanisms for the support of an S&S-directed vision and planning system.

2 Summary of work done in this period

The major visual perception capabilities relevant to security and surveillance that we addressed in this research program are the detection of human subjects, the tracking of their motion and the re-detection of specific, previously monitored, subjects. Research was performed this year towards this end, and progress has been made in the following critical areas:

1. The development and testing of automatic methods of extracting visually salient image parts for the detection of humans.

2. The development and testing of parts-based recognition algorithms for the detection of humans.

3. The study and development of automatic methods of tracking and re-detection of moving human subjects.

4. Research and development support.

The work performed in each of these areas is summarized in this section. An extended technical presentation of the research on visually salient parts extraction for the detection of humans and the application of parts extraction for the task of human face detection can be found in Section 3.

2.1 Visually salient parts extraction for the detection of humans.

One area where interesting progress was made is in use of parts-based descriptions for the recognition of human figures. This work carries forward a project started on a prior ARPA contract. The intent of this work has been to find a method of effectively representing the way the image mass is distributed in a figure. For example, in an intensity image of a person's silhouette, the figure has concentrations of image mass at its limbs and head (if they are showing), and a description of those mass concentrations centers could be used to help discriminate between different shape types. Our previous work had been restricted to intensity images. The study of intensity-based methods has been expanded, and methods based on color and visual texture have been developed.

2.1.1 Extended study of parts extraction in intensity images of human forms.

The study of intensity-based parts extraction has been furthered by more extensive testing, the development of the underlying theory and a detailed presentation.

In order to assess the stability of the representation, the previously developed extraction process was applied to the extraction of salient human parts in extended sequences of intensity images from a video camera. Results have been very good on some types of shapes, such as hands, arms and legs, and mixed on face images. To test the method of parts extraction on human figure recognition, image sequences of actual, articulating human subjects in cluttered scenes were used, and the effectiveness of the parts-extractor for detecting limbs in these sequences was evaluated. Overall, the method appear promising; however, it is sensitive to undersampling of the image prior to Gaussian smoothing and would be more effective in some cases if color information were utilized in addition to intensity.

The theory behind the basic method has been developed more completely, and an extensive comparison with competing techniques was done. This included a mathematical analysis of different multiscale local contrast features with regard to their usefulness in locating regions of local contrast in image position and scale (the salient intensity-based parts). A write-up of this study has been completed and published in the Proceedings of the 1994 ARPA Image Understanding Workshop. The paper contains a presentation of our general approach to people recognition, details of the visual parts analysis (appropriate-scale ridge detection), face feature extraction and face detection; it has been included in this report as Section 3.

2.1.2 Development of a color parts extraction process.

The basic technique has been extended to make use of color data which may provide multiple, partially-independent sources of information for the extraction process and also may be more stable against shading and lighting variations. An initial design and implementation of a figure extractor and shape analyzer based on color image data has been completed. A single test was performed on a full-body color image of a person. All limbs and torso were extracted from the background as separate body parts, with approximate, but reasonable proportions estimated. A description of the process can be found in Section 3.4. More testing and development is required to fully understand how to utilize color and assess its effectiveness.

2.1.3 Development and testing of a texture and shading feature extraction process

As discussed above, the intensity-based parts extractor has mixed results when applied to the complex shading of the human face, an important object to detect in this task domain. To achieve illumination- and view-insensitive face detection, we defind a new parts-based recognition process in terms of oriented texture fields. The parts in this new scheme involved the local maxima, after Gaussian smoothing, of the magnitude of oriented image variation. This new process was implemented and tested on video sequences of moving a face under different lighting conditions. It was found to be reasonably robust with respect to lighting changes on smooth surfaces (e.g., faces) and movement of the surface relative to the camera. A description of the process and examples of its application can be found in Section 3.5.

2.1.4 Design of accelerator for visually salient parts extraction

The parts extraction algorithms for each modality discussed (intensity, color and texture) are based on separable convolutions and local non-linear operations. Thus, the overall process is amenable to high-speed implementation. An initial design of such an implementation was completed and analyzed with respect to field-programmable gate array realization.

2.2 Parts-based recognition algorithms for the detection of humans

Once stable, informative parts are extracted from the image, various methods can be applied to the modeling of human likeness and the detection of human subjects. A basic face detector was developed and tested using the oriented-texture parts; other algorithms were studied for the task of recognizing articulating human figures.

2.2.1 Development and testing of face detection and tracking process

A generic human face detector was developed, based on the illumination- and view-insensitive extracted face parts. We completed the design and implementation of the matcher and a face model for the face-finding task. The face-finder was tested on a 20-frame video sequence containing a head undergoing a range of 3-D rotations. In 16 out of 20 frames the system correctly detected the face and correctly labeled all the major face parts.

A first pass was then made at developing a method for filtering out bad matches in such video sequences using a face motion smoothness criterion. The key idea is that bad matches between our generic face model and the raw data would, with high probability, exhibit very large physical motions away from the positions of correctly matched faces in the video sequence. This approach allowed us to correctly remove the bad matches and interpolate a smooth motion trajectory for the face for all of the video sequence.

Our face detection technique was tested on a broader set of imagery, including over a hundred images of a cluttered scene; it exhibited a 7 percent false positive rate and a 35 percent false negative rate. A description and example of the face detection was published in the 1994 ARPA Image Understanding Workshop proceedings and can be found in Section 3.5 of this report.

2.2.2 Additional face recognition studies

We also continued our fundamental study of image features for use in face detection. The starting point was to consider the face as a smooth surface patch of possibly variable pigmentation.

The brightness contrast of a projected surface patch, relative to its surrounding neighborhood, can vary with lighting and background. However, it may tend to be correlated with the contrast at other points on the same object. Hence relative contrast may be a more stable feature than brightness or brightness contrast alone. This concept can be generalized to all aspects of brightness variation (shading). It is important to note that this is not the same problem as model independent understanding of the 3-D shape from shading problem. Even though much progress has been made on this reconstruction problem [Oliensis, 1990], it remains problematic as it generally requires assumptions about the viewing context and integration across a single, contiguous, Lambertian surface of constant reflectance. By treating shading as a problem of matching to known object surfaces, rather than model-independent understanding, one may be able to dispense with many of the assumptions made about the viewing conditions.

For example, given the relative positions of a set of n model points and their known local surface properties, and a set of image measurements from $n-1$ image points corresponding to the first $n-1$ model points, we may be able to predict some of the shading properties of an additional image point brought into correspondence with the nth model point. Therefore, these additional properties can be treated as measurements to match to the model for verification of the correspondence. The set of local measurements and positions of the image points could be useful for matching the image to models in a manner analogous to the model-based invariants developed by Weinshall [Weinshall, 1993], a method of utilizing the geometry of the projection of non-coplanar object points. The process of developing indexing functions from model-based invariants discussed in [Weinshall, 1993] could also be applicable.

The use of shading-based invariants could be critical for the recognition of many classes of objects having predominantly smooth surfaces, including human faces. The key role of shading information in face identification by humans is demonstrated in [Bruce et al, 1993]. The role of shading understanding is apparently to match properties of the pigmentation and 3-D shape of the face to the image data. Given that model-independent reconstruction of these properties is difficult, model-based approaches such as our proposed shading invariants could be crucial.

2.2.3 Investigation of detection of articulating human figures

A natural extension of our detection studies of practical importance in this task domain is the detection of articulating objects, such as full human figures in motion. To this end, we reviewed and began developing matching algorithms for articulating objects.

We have also studied issues involved in actually representing human figures in support of recognition. The human model must incorporate two general types of parameters: (1) Static parameters, which are fixed (for a given individual), steady-state values, describing such parameters as limb lengths and nominal relative positions between body parts; and (2) Dynamic parameters which describe the position of the body parts at any given time. Dynamic parameters may encode an individual pose, deliberate gestures, or other body movements.

Some requirements we impose on the human figure model are that it must: accept partial information and information at multiple scales; update parameters in the absence of sensory information (e.g. for occluded parts); represent uncertainty in the estimation of the parameters; propagate uncertainty over time; and model realistic human motion by inherent constraints (e.g. no self intersection) and limitations on lengths, joint angles, and velocities.

To accomplish these objectives in the representation, we worked on a layered modeling system which includes a constraints layer, a dynamic model, and a kinematic model. All access to the model (updates and queries) is handled through the constraint layer.

2.2.4 Theoretical support for recognition algorithm development and assessment

In our work on human detection and tracking, we use algorithms that match collections of image features against collections of stored model features searching for correspondences that optimize some measure of likelihood. The robustness of these matching algorithms under conditions of occlusion, noise, and clutter depends heavily on the statistical distribution of characteristic configurations of image features. Although there is little hope of specifying these distributions in full detail across a broad range of viewing situations, it is still useful to study their properties (both abstractly and empirically) in order to gain understanding of essential tradeoffs and improve the reliability of the algorithms.

During this year, work was begun on analyzing the combinatorial properties of a class of matching algorithms and characterizing how the reliability of a match might vary with the size of the feature sets, the number of relations measured, and the degree of independence of those measurements. Preliminary results suggest that, under reasonable assumptions, algorithms in this class can achieve very low error rates at acceptable levels of computational complexity, but that they also exhibit a high degree of sensitivity to key parameters in the ranges of interest (i.e., from several to several dozen features and measured relations) and further highlight the need for independent measurement types when attempting to match non-rigid objects robustly.

2.2.5 Investigation of template-based methods for realtime recognition

We have developed some new ideas about how to approach the recognition problem based on the use of template-based correlation as feature generators. Early results suggest that the technique may be useful for rigid object pose estimation and object classification. We are also working on techniques to use the above template-based ideas to increased the robustness of our correlation-based tracking system. The new techniques for tracking and object pose recovery have been prototyped on our real-time vision hardware. These involve use of pre-stored templates taken from sign of Laplacian of Gaussian-filtered images. This work was originally done in support of an augmented reality project for the Navy; however, we have begun exploring the potential use of that technology in the security domain as a means for recognizing and discriminating prototypical human parts. Since the basic idea is best for rigid, or semi-rigid parts, this idea may be useful for tracking human heads. If applied to the detection and localization of full articulating figures, the templates used must be restricted and scaled to the size of locally-rigid features of the body.

We consider this as part of our theoretical studies that have continued to focus on developing local image features useful for doing human detection and recognition. During this year, we have explored a range of possibilities and are now attempting to narrow the field to techniques that are both robust and computationally feasible for real-time implementation.

2.3 Human motion tracking and activity monitoring

Humans in motion may be detected as things moving in a certain way. In addition, the act of moving about creates a monitoring problem that may require search and re-detection of a specific, previously viewed human subject. These aspects of human detection imply the following work, which has begun this year.

2.3.1 Human motion tracking

On human/non-human discrimination, a literature review on vision-based human motion tracking and human activity monitoring has been carried out. Development was also started on software support for human motion tracking research. We developed the first stage of a window-based graphical interface to display image sequences and moving articulated models, and to visualize recovered three-dimensional models. Routines were developed for extending the 2-D articulated part tracking used for parts-based recognition to the tracking of 3-D moving parts. A graphics system was completed for displaying three-dimensional models in two (front and side) views, allowing for user manipulation of the whole model or individual parts.

2.3.2 Color histograms for simple attentional and person re-recognition mechanisms

We began investigating the use of color histograms as a simple attention mechanism and to quickly re-recognize a person. We also began to build software to test and experiment with color histogram-based algorithms. The color histogram routines for tracking a previously identified target were developed.

2.4 Research and development support

Additional work was performed in support of the research and development effort. It was directed towards hardware and software support, assessment of requirements and current technology, and meeting with project supervisors and research peers.

2.4.1 Hardware and software support

Software was developed to capture image sequences for human tracking and face recognition experiments. The programs capture sequences of intensity or color data, at specified image sizes and capture rates, and display image sequences for inspection.

On development of an experimental surveillance system, we conducted a product survey of motorized zoom lenses. Also, a new security camera was acquired from a collaborating partner in the security industry. This camera allows high speed pan and tilt movement as well as control over zoom, focus, and lens aperture. The real-time control interface for this sensor will prove useful in our research on detecting and tracking intruders.

2.4.2 Requirements and technical assessment

Discussions were held with contacts at several major security and surveillance suppliers to assess the current state of available technology and currently perceived needs of the industry. The purpose of these discussions was to identify key areas to focus our research.

2.4.3 Meetings

We attended an ARPA workshop on Advanced Vision Processors (AVIS) organized by Barbara Yoon. This meeting brought together users and developers of advanced vision processing systems to discuss directions for the future.

On March 29, 1994, Lauretta Williams, the TEC COR for this contract, and Oscar Firschein of ARPA visited Teleos for a project kick-off meeting. Presentations were given describing our objectives for the research program and our current progress. We also had a productive interchange of ideas on the relation of our work to other ARPA programs and were invited to make a contribution to the IU workshop proceedings. This invitation was accepted and resulted in a paper which is included as Section 3 of this report.

3 Detailed technical presentation

This section is a more detailed technical presentation of the research on visually salient parts extraction for the detection of humans and experiments in human face detection via parts-based recognition methods. This work has recently been published in the 1994 ARPA Image Understanding Workshop. In this presentation, the concept of image representation in terms of local centers is motivated, and computational models of the concept are compared. A particular model, the *appropriate-scale ridge*, is developed and demonstrated. In this model, local centers are defined as smoothed image extrema that are also maximal, with respect to scale, in the magnitude of the second spatial derivatives. The basic ideas are extended to color and texture data, and the texture features are demonstrated via face detection.

3.1 Introduction

A major goal of the research presented here is the recognition of natural objects given natural viewing conditions. An important example of this problem is the visual detection, identification and understanding of people. The recognition of people requires the extraction of visual information that tends to be stable with respect to such transformations as limb articulation and variations in clothing. Given this, image properties that are potentially useful for recognition include aspects of the proportions and geometric arrangement of visually distinct object parts. Thus, our basic approach to recognition is to extract key visual parts, model objects as relational graph descriptions over these parts, and recognize the objects by matching images to the graph descriptions. Since graph relations can be arbitrary, the relational graph formalism is sufficiently general to include approaches to rigid object matching such as alignment-based methods [Huttenlocher and Ullman, 1987], as well as recognition strategies for more variable configurations of parts. By analyzing partial graph matches that can be efficiently searched and yet provide strong indications of object presence, parts-based and graph-based recognition can be made fast and effective [Burns and Rosenschein, 1993].

Parts-based recognition has been proposed at least as early as [Marr and Nishihara, 1978], and is in contrast with current image-based methods for people detection and recognition [Beyer, 1993, Bichsel and Pentland, 1994]. To detect people in all possible articulations, views and lighting, image-based approaches would require a large collection of reference images. In addition, Bichsel and Pentland observe that, since the set of images being sampled is highly non-convex, its linear approximation (e.g., eigenimages) has very limited effectiveness for reliable detection. Given this, parts-based recognition appears more promising.

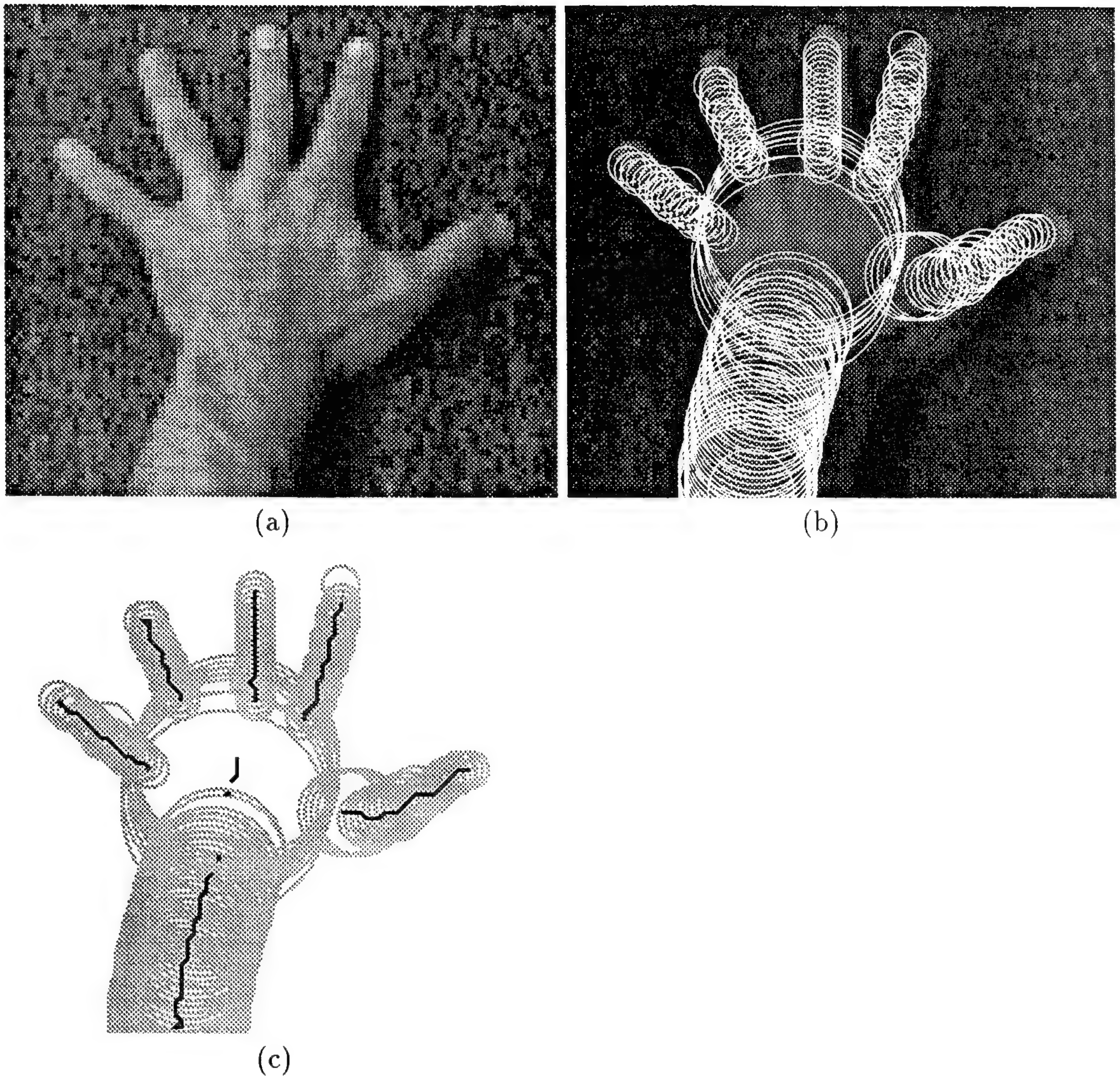
This paper presents a computational model of detectable visual parts that are a potential foundation for parts-based recognition. In general, such parts should be visually stable and informative in the sense that lower order relations over them provide discriminating properties for natural objects.

Given these objectives, it appears difficult to formulate a model of appropriate visual parts in terms of either local image edges or 2-D connected components. A set of extracted edges generally provides too low a level of representation for recognition: single edges usually represent little information; matching large numbers of edges can be very slow; and considerable effort may be required to produce more compact representations from a set of edges. Edge level representations may be useful in situations where a single, rigid reference frame can be used to evaluate the registration of fragmented data [Huttenlocher and Ullman, 1987]; however, such a context is often not available in natural object recognition.

Image connected components generated by pixel classification (image segmentation) tend to have problems with stability: small changes in view or lighting can produce dramatic changes in the topology of the connections. This in turn produces dramatic changes in the size, shape and position of major components. In addition, complex assemblages of parts are often represented by a single component, requiring potentially expensive shape analysis to extract the salient parts.

Instead of edges and 2-D connected components, we define salient image parts in terms of *local centers*: visually compact regions that have significant internal-external value contrast in any of various image measurements, such as brightness, color, texture, depth or motion. Given this model, local centers are extracted that optimize internal-external contrast with respect to position and scale of the center; the local centers can then be organized into basic visual parts via simple geometric rules. Local centers need not be mutually exclusive: centers of different sizes, representing significant structures at different scales, can be associated with the same location in the image.

The advantages of this scheme over an edges-only model are that (1) the basic unit of representation already encodes something about the position,



(a)

(b)

(c)

Figure 1: An example of a natural object readily represented as a set of *local centers*, visually compact regions that have significant internal-external contrast. The local centers shown have been extracted using the appropriate-scale ridge model discussed in the text: (a) image, (b) scale and position of detected centers represented by radius and center of circles, and (c) centers linked into parts.

size and internal characteristics of a visually distinct part of the object, and (2) the extracted centers potentially require less organizational processing and selection than edges to be useful for recognition. The advantage of this scheme over 2-D connected components is that the basic units are not as complex and unstable. Figures 1-4 show examples of objects that can be readily and fruitfully represented as sets of local centers of brightness value (Figures 1 and 3) and local centers of texture orientation and magnitude (Figure 4). The local center representation need not be used to the exclusion of extracted edges, which might further localize the boundaries between regions of complex shape. However, we believe that explicit edge information is not always reliable or required and should often be organized with respect to the extracted local centers.

The rest of the report surveys models of representation related to local centers, develops a definition for a particularly useful model, the *appropriate-scale ridge*, and demonstrates local center extraction on stable representation and recognition.

3.2 Models of multi-scale local centers

This section presents computational models of local centers in terms of intensity contrast; later sections will expand the discussion to include color and texture.

The basic models can be characterized as medial axis transforms, Laplacian of Gaussian peaks and ridges, intensity peaks and ridges, and what is referred to here as *appropriate-scale ridges*. The latter model combines the concept of intensity ridge with local contrast maximization across scale to effectively localize in both image position and scale.

3.2.1 Medial axis transforms

The basic idea of reducing complex shapes to a set of local centers has been proposed at least as early as the medial axis transform of [Blum, 1973, Blum and Nagel, 78]. The transform reduces a shape to the set of circles of varying position and scale that are multi-tangent with the shape. Since these circles lie along the axes of shape parts, they can be used as a basis of parts extraction and representation. However, the basic transform is very sensitive to small perturbations of the shape boundary, thus multi-resolution versions have been developed [Dill et al., 1987, Pizer et al., 1987]. Unfortunately, the basic idea still requires that an intact region be extracted prior to analysis. Also, the extraction of shape parts is not a function of the local contrast of the parts themselves. Both of these properties limit the usefulness of the idea as a general method of representing visually salient image structures.

3.2.2 Peaks and ridges of LoG

One way to more directly associate local centers with places of local maximum internal-external contrast is to define them as the peaks and ridges of the image convolved with the Laplacian of the Gaussian (LoG) [Nishihara, 1988][1]. The peaks of the difference of low-pass filter output is essentially the same idea [Crowley and Parker, 1984]. This model is also related to the bar detector designed by Canny. In [Canny, 1986], the image is convolved with a mask that has a cross-section, well approximated by the second spatial derivative of the Gaussian. Peaks in the magnitude of the output are then extracted.

In the LoG model, an image structure is represented by a collection of LoG peaks and ridges at different scales. The LoG operator at a given scale and image position is simply the difference between the weighted average of a central zone of width related to the scale and the weighted average of a surrounding zone. Thus, peaks in the output clearly represent contrast maxima with respect to image position. Given the Gaussian filter, the weights define a smooth envelope over each zone, and the output extrema exhibit stability with respect to significant image transformations [Nishihara, 1988]. In addition, the number and positions of the extrema at different scales vary as a function of some interesting shape attributes, such as the number of significant bumps on the shape and the pointedness of the bumps. Thus, the set of peaks and ridges of the LoG output generated at multiple scales may provide a useful representation for shape discrimination.

One property of the LoG peaks and ridges is their sensitivity to intensity edges or shoulders. Basically, local extrema appear at image positions where the center region is at one side of an abrupt intensity change and a maximal portion of the surrounding region is on the other side. This edge response

[1] For this discussion, we will refer to both peaks and pits as *peaks*, and ridges and valleys as *ridges*.

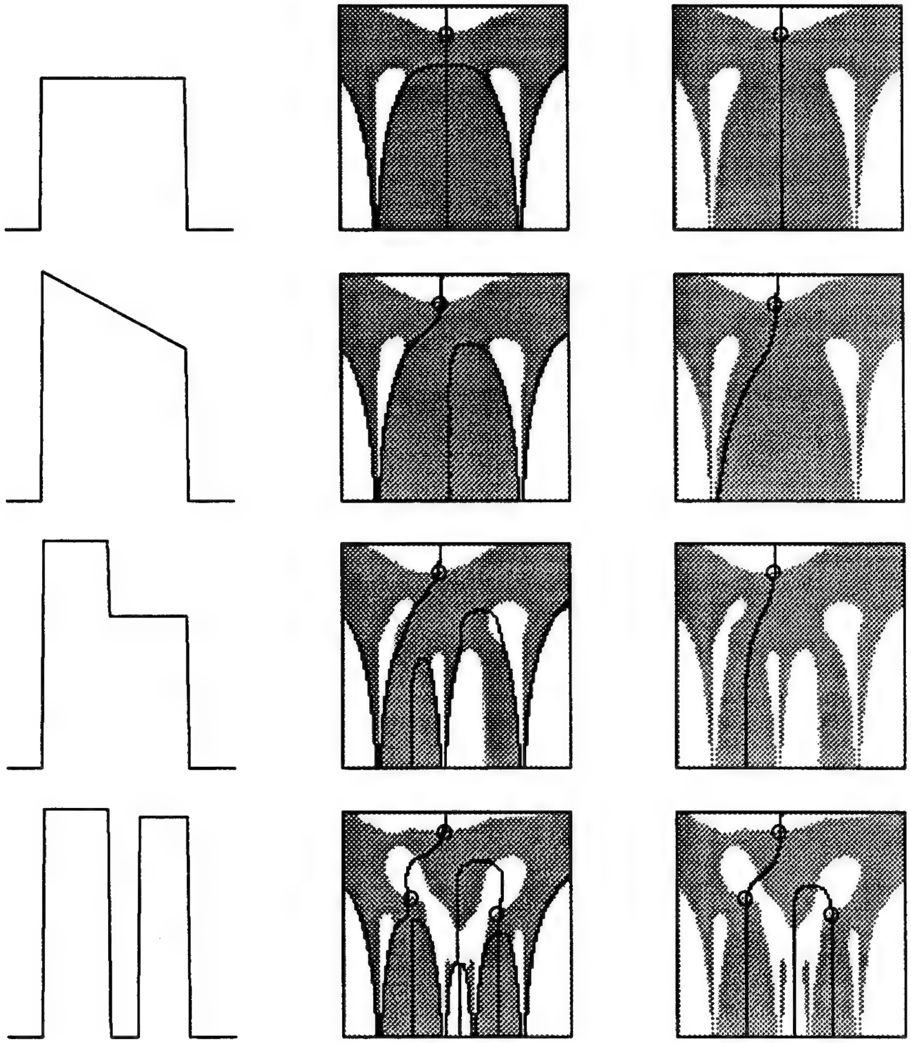


Figure 2: Scale-space analysis of G_{xx} versus appropriate-scale peaks (see text for discussion). First column: 1D images (x by intensity); second column: x by $\log \sigma$ plots of $G_{xx\sigma}$ sign bits (white/gray = $+/-$) with G_{xxx} zero-crossings (black); third column: same sign bit plot with G_x zero-crossings (black). Circles mark position and scale of desired local centers.

happens at every scale the edge is present, can be of high magnitude, and has a position that shifts as a function of the scale.

The objective of the present study is to extract local centers that are defined as compact *regions* of significant internal-external contrast. Given this objective, an extracted local center should represent the overall positions and extents of an object and its distinct parts, and thus provide a compact, hierarchical decomposition of complex object shapes. This objective could be satisfied by an LoG peak-and-ridge model if the edge responses could be suppressed or labeled. This has been proposed in [Crowley and Parker, 1984]; however, it is not generally practical. The LoG peaks with respect to image position (x, y) can be due to whole regions (local centers) or edges. The difference between the two types is that the former peaks are also local extrema with respect to the Gaussian scale parameter σ, while the latter are not. Thus, the suppression of edge response is essentially equivalent to finding LoG local extrema with respect to σ as well as image position (x, y). Given most practical situations, only a discrete sampling of σ is available and true peak detection is difficult.

This can be understood by considering the equivalent 1D case: detecting peaks in $G_{xx}(\sigma) * I$ with respect to (x, σ), where $G_{xx}(\sigma) * I$ is the convolution of 1D image I with the second derivative of the Gaussian at scale σ. These peaks are at points (x, σ), where zero-crossings of $G_{xxx}(\sigma) * I$ and $G_{xx\sigma}(\sigma) * I$ coincide. In order to select only true peaks with respect to both σ and x, the zero-crossings of $G_{xxx}(\sigma) * I$ and $G_{xx\sigma}(\sigma) * I$ must be far enough apart from each other in the neighborhood of edges to not appear coincidental given discrete sampling methods. The examples in the second column of Figure 2 show that this misleading coincidence at edges may often be hard to avoid in practice. The first column shows 1D images (x by intensity) and the second column shows the scale-space plots [Witkin, 1983] of $G_{xxx}(\sigma) * I$ and $G_{xx\sigma}(\sigma) * I$ (x by $\log \sigma$) in response to each image. The gray represents areas of negative $G_{xx\sigma}(\sigma) * I$, while white is positive (the borders of the two areas are the zero-crossings). The thick black lines are zero-crossings of $G_{xxx}(\sigma) * I$. True peaks in $G_{xx}(\sigma) * I$ with respect to (x, σ) should be near the desired local centers, represented as circles in the plots.

As can be seen, there are zero-crossings of $G_{xxx}(\sigma) * I$ and $G_{xx\sigma}(\sigma) * I$ that *appear* coincidental in the finite sampling of σ, are not near the true peaks and are instead associated with edges. This occurs even with the one hundred samples of σ used in this example, which is generally not practical. Tracking the zero-crossing curves through scale-space and selecting points of extreme $G_{xx}(\sigma) * I$ with respect to a parametrization of the curves might accomplish the desired effect, though this would seem to be a complex process, sensitive to sampling and noise. It is likely to be even more difficult to analyze the zero-crossing *surfaces* in (x, y, σ).

Given the objective of the current study, and the difficulties of using a straightforward LoG peak-and-ridge scheme as a model of local centers, an alternative method has been developed. However, the patterns of LoG peaks across scales still may be a useful means of uniquely characterizing shapes for their discrimination, once they have been extracted via a local center process.

3.2.3 Peaks and ridges of intensity

In [Subirana-Vilanova and Sung, 1992], the problem of an edge response is avoided by a more complex set of operators. The bar detection operator of [Canny, 1986] is "split" into two complex assymetric operators which are combined by selecting the minimum magnitude response of the two at every point. The resulting non-linear operator is applied everywhere in the image at different directions and scales, and the maximum output at each point is selected. The resulting output surface is then analyzed for skeletons, much like systems of peaks and ridges. The method does appear to avoid edge responses; however, it is not clear why the process prior to the peak and ridge extraction is necessary. Directly detecting peaks and ridges in smoothed images may work as well, or better. The two contributions to local center analysis made in [Subirana-Vilanova and Sung, 1992] are that the scale of the underlying image structure is also recovered and the operator has been extended to color images. We believe that these two aspects of the problem can be handled without requiring such a complicated system. This is discussed in the next sections.

A potentially much simpler way of insuring that the *centers* of visually distinct concentrations in value are detected, instead of the edges, is to use directly the peaks and ridges of smoothed intensity. They are always associated with the centers of some relatively brighter or darker region at a given scale [Gauch and Pizer, 1993]. By extracting intensity extrema over a range of scales, a compact description of the image related to local centers may be possible. This basic idea has been studied in [Gauch and Pizer, 1993], with related ideas developed in [Koenderink, 1984 and Lindeberg, 1993]. The concept is also consistent with the observation in [Morrone and Owens, 1987] that edge-like and bar-like (i.e., ridge or local center) image features can be distinguished by the relative magnitude of the outputs of odd versus even operators (phase). They studied 1D operators that closely resembled G_x (odd) and G_{xx} (even) at a fixed scale. The centers of the bars were at points where the odd function is zero, and not the even function, and vice versa for centers of edges. This is equivalent to detecting ridges at the zero-crossings of G_x. Results in [Gauch and Pizer, 1993, and Lindeberg, 1993] show that detected intensity peaks and ridges correspond to intensity regions and curvilinear features.

In [Gauch and Pizer, 1993], extrema detected at one scale are related to those at other scales via a scale-space tracking process, though their methods seem complex and tentative. Given tracked extrema, the scale of the underlying image structure has been either defined as the scale at which the associated extrema are annihilated [Lindeberg, 1993] or defined as a *set* of positions and scales [Gauch and Pizer, 1993]. For intensity ridges, the latter output is in the form of a 2-D sheet in scale-space.

It seems problematic to define scale in terms of the annihilation point, since this event is a function of structures external to the local center or region being tracked. For example, the annihilation will occur at a much smaller scale if the region is between two near and large regions than if it is between two far and small ones. Likewise, the complex 2-D sheet representations seem unwieldy and only indirectly related to the underlying properties of width, position and contrast of a given region. Given the complexity and indirection of methods based strictly on intensity extrema, it is useful to consider other possibilities.

3.2.4 Appropriate-scale ridges

We have developed a model of the local center concept called the *appropriate-scale ridge*, which emphasizes the best of both the LoG and intensity ridge approaches. Like the LoG model, internal-external contrast determines the scale and saliency of a local center, and, like the intensity model, the position of a local center is constrained to being at an intensity peak or ridge.

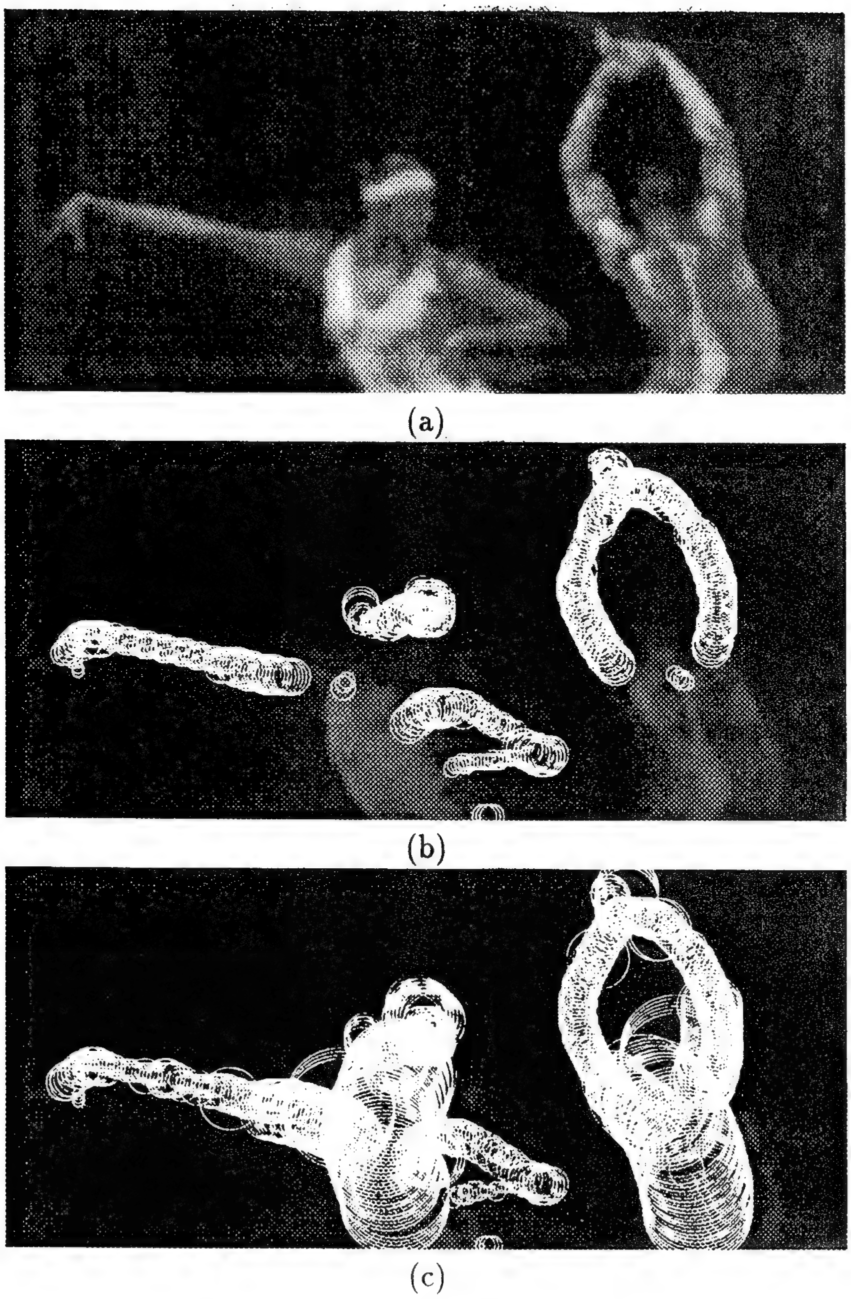
This approach is best understood in the 1D intensity case. The third column of Figure 2 shows some examples for 1D images of appropriate-scale ridge points. First, consider the example in the first row: a 1D box-shaped region of some width that is bright relative to its surround. Over a range of scales σ, the position x_{max} of the maximum of $G(\sigma) * I$ with respect to image position x will correspond to the centroid of the region, where $G(\sigma) * I$ is the convolution of image I with the Gaussian at scale σ. Thus, this extremum is a good model of the associated local center's position. The local center is visually significant if its internal-external contrast is high enough. In 1D, a useful center-surround contrast operator at a given scale σ is simply $G_{xx}(\sigma) * I$, where $G_{xx}(\sigma)$ is the second derivative of the Gaussian with respect to x. Thus the local center contrast is high if the magnitude of $G_{xx}(\sigma) * I$ is high at x_{max}. In addition, the scale σ that best corresponds to the width of the underlying region is the one for which the magnitude of $G_{xx}(\sigma) * I$ at x_{max} is maximal.

In summary, this implies the following useful model of a 1D local center. A point (x, σ) is an appropriate-scale peak if it is:

1. At a local extremum, with respect to image position x, of $G(\sigma) * I$, and

2. At a local extremum, with respect to scale σ, of $G_{xx}(\sigma) * I$. (The senses should also be compatible: a minimum of $G_{xx}(\sigma) * I$ if $G(\sigma) * I$ is at a maximum, and vice versa.)

Points that satisfy the above conditions are at intersections of the zero-crossings of $G_x(\sigma) * I$ and $G_{xx\sigma}(\sigma) * I$ in scale-space (x, σ). Specifically, they are at the intersections where the sign change for $G_x(\sigma) * I$ in the direction of positive x is the opposite as the sign change for $G_{xx\sigma}(\sigma) * I$ in the direction of positive σ. The zero-crossings for each image in the first column have been plotted (x by $\log \sigma$) in the last column of Figure 2. The gray represents areas of negative $G_{xx\sigma}(\sigma) * I$, while white is positive (the borders of the two areas are the zero-crossings). The thick black lines are zero-crossings of $G_x(\sigma) * I$. The circles represent the position and scale (x, σ) of desired local centers in the image. As can be seen, the examples include 1D regions with asymmetric bumps and gaps of significant magnitude. In each case, zero-crossings of compatible sign change coincide at points that are near the centroid and width of the actual regions of significant contrast (i.e., no edge effects); elsewhere, the compatible zero-crossings are well separated. Thus in 1D, the appropriate-scale peak model appears to correspond to our concept of a local center.



(a)

(b)

(c)

Figure 3: Appropriate-scale ridges of human figures: (a) image, (b) ridge output from four scales, and (c) ridge output from five scales.

3.3 Appropriate-scale ridges in 2-D

A version for 2-D images (x, y) can be readily defined in terms of the second directional derivatives of $G(\sigma) * I$, or $G_{tt}(\sigma) * I$ for image position parameter t, varying in some image direction θ with respect to the x axis. The direction of the maximum magnitude second directional derivative at (x, y, σ) corresponds to the direction oi maximum internal-external contrast at this point and scale; the positional parameter in this direction will be referred to as $t_{\mathbf{max}}$. Given this, the following model is used. An image point and scale (x, y, σ) is an appropriate-scale ridge point if it is:

1. At an extremum in $G(\sigma) * I$, with respect to t_{max}, and

2. At an extremum, with respect to σ, in $G_{t_{\text{max}} t_{\text{max}}}(\sigma) * I$, the local contrast at (x, y, σ). (Again, the senses should also be compatible: a minimum of $G_{t_{\text{max}} t_{\text{max}}}(\sigma) * I$ if $G(\sigma) * I$ is at a maximum, and vice versa.)

In order to compare the local contrast measured at different scales σ, the volume defined by the operator surface in its negative (center) and positive (surround) parts must be made constant with respect to σ. This can be achieved by multiplying the operator by σ^2.

For the discrete implementation of this model, σ is sampled exponentially [Koenderink, 1984]. For each sampled scale σ_i, 2-D image ridge points are detected using rule (1) above, and a ridge point (x, y, σ_i) is selected if its local contrast exceeds the local contrast at (x, y, σ_{i+1}) and (x, y, σ_{i-1}).

Figures 1 and 3 show the results of applying this process to images of natural objects with parts of various positions and widths. In each example, a local center is shown as a circle width radius and center corresponding to the local center scale and position. Only the bright centers are shown in each case; the dark centers define the shape of the complementary regions. In Figure 1(b), output is shown from five scales, and a contrast threshold of 10 was used. Note that the widths seems appropriate for the underlying substructures (up to the scale sampling) and the parts are densely sampled along their axes. Figure 1(c) shows the extraction of salient object parts by simply linking neighboring centers into chains. Figure 3 shows the results on human figures. In Figure 3(b), output at four scales is shown (from finger to arm widths), and Figure 3(c) shows 5 scales (from finger to trunk width). Note that the operator avoided producing edge responses, and the overall shapes of these complexly shaded figures are reasonably represented by the general configuration and widths of the local centers.

3.4 Color model

It is useful to expand the concept of an appropriate-scale intensity ridge to multimodal input. The use of multiple channels of input increases the likelihood that a given local center of contrast is associated with a figure that should be considered distinct from its surround: the more ways in which the internal data differs from the external, the more likely the internal region is of a different material and scene object. Color provides such multiple channels. Also, given typical scene lighting, color components such as hue are relatively insensitive to internal variations in the object projection due to shading. Thus with color contrast, the overall shape of the figure outline (e.g., someone's shirt or pants) is often emphasized more than the details (e.g., the cloth folds and wrinkles).

A color image is a mapping from two parameters (x, y) to a vector of color components (u, v, w). Since it is a parametrization of a 2-D manifold in the color space, there are no "peaks" or "ridges" in the sense defined for intensity images. However, a simple extension of the ridge or peak concept can be used. Consider a 1D image, $I : x \mapsto (u, v, w)$. Regions of relatively constant color in the image correspond to sections of the manifold where the parametrization "slows down"; in other words, the arclength of the parametrization $s(t)$ is relatively small. Consistent with this, image color edges are places where $s(t)$ is relatively large. Thus, we can construct an analogous concept to the intensity case by defining a color peak as a point of local minimum arclength $s(t)$, and an appropriate-scale one as a peak that is also a maximum with respect to smoothing scale σ in the second derivative of $s(t)$ with respect to t.

For a 2-D image, $I : (x, y) \mapsto (u, v, w)$, there is an analogous concept of appropriate-scale color ridge. This has been developed and implemented. It has been given a tentative testing and evaluation on a full-body color image of a person (not shown here). All limbs and the torso were extracted from the background, each as a separate body part, with approximate, but reasonable estimates for the part proportions.

3.5 Texture and face detection

Sometimes a figure's internal shading variation is also important for parts extraction. For example, salient features of an object surface, such as a nose on a face, often appear as a patch of image shading or texture. To recognize faces, it is useful to extract local centers that correspond to concentrations of certain shading or texture attributes. For this reason, the ideas of Gaussian smoothing and peak detection have also been applied to local centers of texture and shading properties; however, currently only a single-scale version has been implemented and tested.

Many aspects of a local intensity patch vary considerably as the light source or camera angle changes. One aspect appears relatively stable and was used as the basis for face representation: the dominant orientation of the intensity variation in an image patch relative to the face's vertical axis. The magnitude of the variation of the intensity in a patch can be different for different directions of measurement. The variation can be measured in terms of the first, second or other derivatives of the image function. By defining the orientation of a texture or shading patch as the dominant orientation of the intensity variation, and by measuring the orientation with respect to the object's reference frame, we have a feature that is relatively stable with respect to lighting and camera change. The salient features of the human face (e.g., the nose, mouth, eye and cheek regions) generally have an expected, dominant orientations, thus local centers of texture orientation and magnitude could be useful features to extract for face detection.

A type of texture feature was developed that is related to this idea, though others could be used. It is based on the fact that smoothing a texture field suppresses the resulting texture magnitude in areas without coherent texture orientation, at the given scale, while keeping the magnitude of coherent textures constant. The effect is to produce peaks in magnitude at points with texture that is relatively coherent or contrasting in orientation from the surround. The specific model is defined as the following steps:

- An oriented texture field (dx, dy) is computed by detecting ridges in the LoG output at some scale and using the ridge direction as the texture orientation (Modulo 180 degrees: ridges of similar orientation but different sign are considered parallel. This is accomplished by multiplying the vector angle by two.)

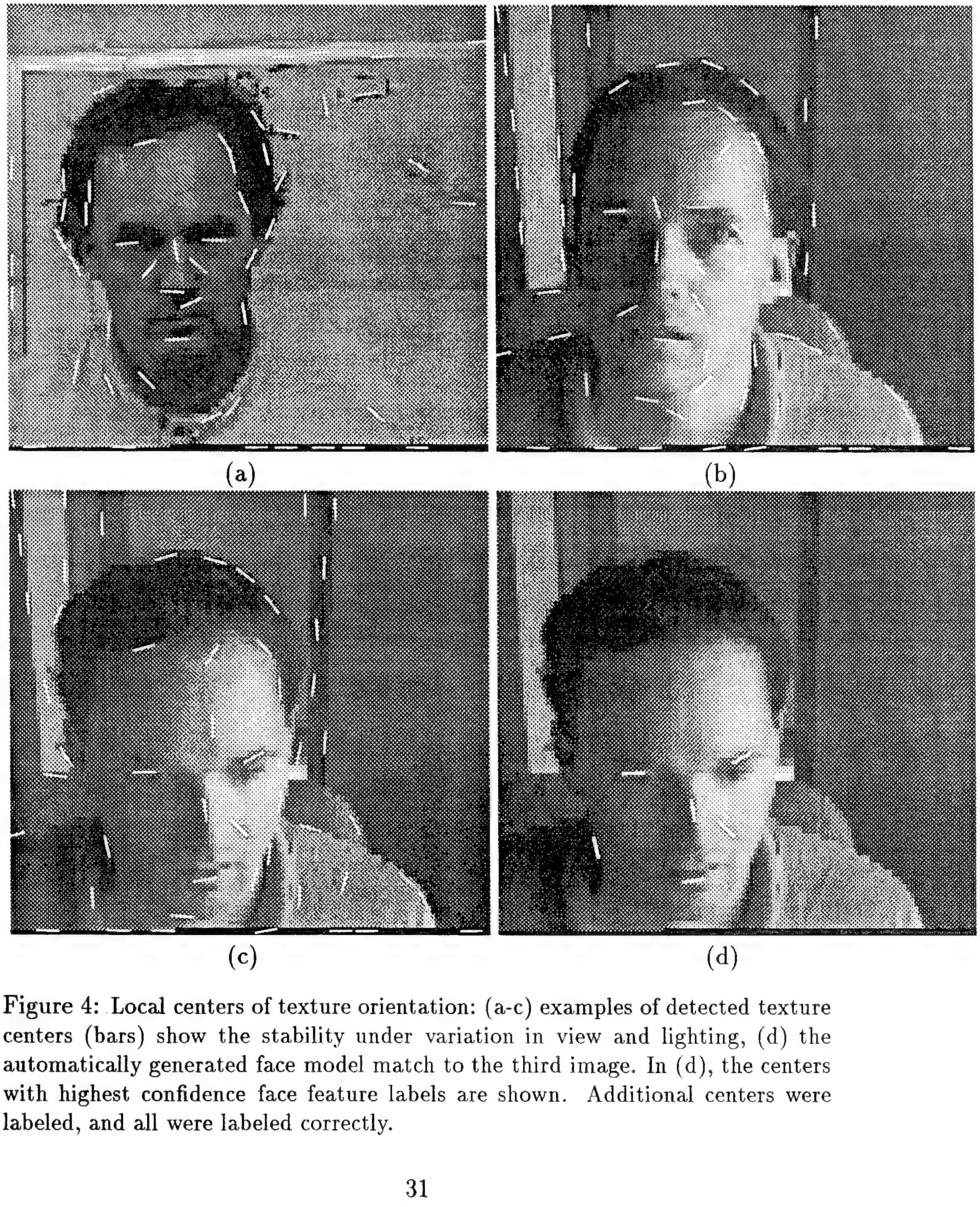


Figure 4: Local centers of texture orientation: (a-c) examples of detected texture centers (bars) show the stability under variation in view and lighting, (d) the automatically generated face model match to the third image. In (d), the centers with highest confidence face feature labels are shown. Additional centers were labeled, and all were labeled correctly.

- The texture field is then smoothed by convolving the individual components dx and dy by a Gaussian.
- The peaks in the resulting texture magnitude $|dx, dy|$ above some threshold are selected as the texture tokens; the orientation of the smoothed field at each peak point is used as the token orientation.

This scheme was tested on a set of face images with different lighting and head positions, holding operating parameters constant. Figures 4(a-c) show example results, where peaks are drawn as bars. In general, the detection and localization seems reasonably stable. The output was evaluated by using it to detect faces. Detection was accomplished by searching for the correct 2-D similarity transformation between each image and a 2-D face model constructed of similar features. Candidate transformations were generated by assigning pairs of model tokens to pairs of image tokens and computing the model-to-image transformation that best aligned them. Each candidate match was evaluated by checking the alignment of the rest of the model tokens to the nearest image tokens. The face detector was tested on a 20 frame video sequence of a 3-D-rotating head. Faces were correctly matched in 16 out of 20 of frames; see for example, Figure 4(d). The incorrect matches were filtered out via smooth motion modeling of the head. With the match score threshold set so that 13 of the correct matches are selected, no false matches are selected in the 20-frame sequence and only 7 out of 100 false matches were accepted in a sequence of 100 frames panning a cluttered scene without faces.

3.6 Summary of detailed technical presentation

In this section, the concept of image representation in terms of local centers is motivated, and computational models of the concept are compared. A particular model, the *appropriate-scale ridge*, is developed and demonstrated. In this model, local centers are defined as smoothed image extrema that are also maximal, with respect to scale, in the magnitude of the second spatial derivatives. The basic ideas are extended to color and texture data, and the texture features are demonstrated via face detection.

Future research includes the continued testing and development of the color and texture versions, and the automatic construction of models of articulating forms such as humans from extracted local centers.

4 Report summary

This Annual Report details progress that Teleos has made in the development of computer vision and visual attention mechanisms for the support of a S&S-directed vision and planning system.

The major visual perception capabilities relevant to security and surveillance that we addressed in this research program are the detection of human subjects, the tracking of their motion and the re-detection of specific, previously monitored subjects. Research was performed this year towards this end, and progress has been made in the following critical areas:

1. The development and testing of automatic methods of extracting visually salient image parts for the detection of humans.

2. The development and testing of parts-based recognition algorithms for the detection of humans.

3. The study and development of automatic methods of tracking and re-detection of moving human subjects.

4. Research and development support.

The work performed in each of these areas is summarized in the report, and an extended technical presentation is given of the research on visually salient parts extraction for the detection of humans and the application of parts extraction for the task of human face detection.

References

[1] D.J. Beyer, "Face recognition under varying pose," MIT. A.I. Memo No. 1461, 1993.

[2] M. Bichsel and A.P. Pentland, "Human face recognition and the face image set topology," *CVGIP: Image Understanding*, 59(2): 254-261, March 1994.

[3] H. Blum, "Biological Shape and visual science, part 1," *J. Theor. Biol.*, 38: 205-287, 1973.

[4] H. Blum and R.N. Nagel, "Shape description using weighted symmetric axis figures," *Patt. Rec.*, 10: 167-180, 1978.

[5] V. Bruce, A. Coombes and R. Richard, "Describing the shapes of faces using surface primitives", *Vision and Image Computing*, 1993.

[6] J.B. Burns and S. J. Rosenschein, "Recognition via Blob Representation and Relational Voting," *Proc. IEEE Conf. on Signals, Systems and Computers*, November 1993.

[7] A.R. Dill, M.D. Levine, P.B. Noble, "Multiple resolution skeletons," *IEEE PAMI*, 9(4): 495-504, July 1987.

[8] J. Canny, "A computational approach to edge detection," IEEE PAMI 8:679-698, 1986.

[9] J.L. Crowley and A.C. Parker, "A Representation for Shape Based on Peaks and Ridges in the Difference of Low-Pass Transform," *IEEE PAMI*, 6(2): 156-169, March 1984.

[10] J.M. Gauch and S.M. Pizer, "Multiresolution analysis of ridges and valleys in grey-scale images," *IEEE PAMI*, 15(6): 635-645, June 1993.

[11] D.P. Huttenlocher and S. Ullman, "Object Recognition using Alignment," *Proc. ICCV*, pp. 102-111, 1987.

[12] J.J. Koenderink, "The structure of images," *Biol. Cyber.*, 50: 363-370, 1984.

[13] T. Lindeberg, "Detecting salient blob-like image structures and their scales," *IJCV*, 11(3): 283-318, 1993.

[14] H.K. Nishihara, "A System for Recognizing the Shape of Printed Words," AAAI Spring Symposium, March 1988.

[15] D. Marr and H.K. Nishihara, "Representation and Recognition of Spatial Organization of Three-dimensional Structures," Proc. of the Royal Soc., B200, p. 269-294, 1978.

[16] M.C. Morrone and R.A. Owens, "Feature detection from local energy," *Patt. Rec. Lett.*, 6: 303-313, 1987.

[17] J. Oliensis, "Uniqueness in Shape from Shading," *International Journal of Computer Vision*, 6(2):75-104, 1991.

[18] S.M. Pizer, W.R. Oliver, S.H. Bloomberg, "Hierarchical shape description via the multiresolution symmetric axis transform," *IEEE PAMI*, 9(4): 505-511, July 1987.

[19] J.B. Subirana-Vilanova, and K.K. Sung, "Perceptual organization without edges," *Proc. IUW*, pp.289-298, Jan. 1992.

[20] D. Weinshall, "Model-based Invariants for 3D Vision," in *Proceedings: IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 695-696, June 1993.

[21] A. Witkin, "Scale-space filtering," *Proc. of the Ninth International Joint Conf. on Artificial Intelligence*, Karlsruhe, West Germany, pp.1019-1022, 1983.